Section II. Technical Notes

The data on doctoral scientists and engineers contained in this report come from the 2001 Survey of Doctorate Recipients (SDR),[1] which is a longitudinal panel survey of individuals who have received their doctorates in the sciences or engineering (S&E). Since the 1970s, this study has been conducted every two years for the National Science Foundation (NSF) and other federal sponsors.[2]

The U.S. Census Bureau conducted the survey for the NSF in 2001. Data collected in the SDR are part of the Scientists and Engineers Statistical Data System (SESTAT) surveys that are sponsored and maintained by NSF. Additional data on education and demographic information in the SDR come from the Survey of Earned Doctorates (SED), an ongoing annual census of research doctorates earned in the United States since 1920, which forms the Doctorate Records File (DRF).

The Sampling Frame and Target Population

The sampling frame for the 2001 SDR was compiled from the DRF to include individuals who:

1. Had earned a doctoral degree from a U.S. college or university in a S&E field[3]
2. Were U.S. citizens or, if non-U.S. citizens, indicated they had plans to remain in the United States after degree award
3. Were under 76 years of age

The 2001 frame consisted of the 1999 SDR sample supplemented with new S&E doctorate graduates who had earned their doctoral degrees since the 1999 survey and who met the conditions listed above. Those who were carried over from 1999 but had attained the age of 76 (or were deceased) were deleted from the frame.

The survey had two additional eligibility criteria for the survey target population. The sampled member must be a resident of the United States and not institutionalized as of the survey reference week (week of April 15, 2001).

Sample Design

In 2001, the SDR sample size was 40,000. The total sample was selected from three groups:

- Old cohort cases with doctoral degrees earned prior to July 1, 1994
- Nearly new cohort cases with doctoral degrees earned between July 1, 1994 and June 30, 1998
- New cohort cases with doctoral degrees earned between July 1, 1998 and June 20, 2000

The goals of the 2001 SDR sample design included the following:

- Reduce the variation in the sampling weights of the old and nearly new cohorts
- Allocate the sample so that the variance of overall population estimates are minimized
- Allocate the sample so that the sampling rate of the new cohort is at least 15 percent higher than that of the old cohort
- Allocate the sample so that the sampling rate of the nearly new cohort is at least 10 percent higher than that of the old cohort
- Adjust the sample allocation if any large stratum receives a disproportionate amount of sample

To ensure that the sampling rate of the new cohort was at least 15 percent higher than that of the old cohort, 4,000 of the total sample was drawn from the new cohort group. The remaining 36,000 sample cases were then divided so that the nearly new cohort would have a 10 percent higher sample allocation than the old cohort.

Using these sample sizes, the sample for the 2001 SDR was selected in two phases. The old and nearly new cohort cases were selected in the first phase and the new cohort cases were selected in the second phase. The sampling was separated into two phases to allow time for more complete sampling data for the new cohort cases.

[1] The discussions presented here are partly from the 2001 Survey of Doctorate Recipients Methodology Report (Census Bureau, 2002).

[2] In 2001, the National Institutes of Health cosponsored the SDR with NSF. In previous rounds, the Department of Energy and the National Endowment for the Humanities co-sponsored the survey.

[3] See Appendix A for a list of the science and engineering fields included in the 2001 SDR sampling frame.

The 2001 SDR sampling frame was constructed from two sources: the 1999 SDR sample and the 1999 and 2000 Survey of Earned Doctorates (SED). The 1999 SDR sample provided information on S&E doctoral degrees earned prior to the 1998-99 academic year. The 1999 and 2001 SED provided information on doctoral degrees earned during the 1998-99 and 1999-2000 academic years.

The 1999 SDR sample cases could be considered ineligible for the 2001 SDR sampling frame for a variety of reasons, e.g., the cases from the 1999 sample who are age 76 and over, found to be deceased, never earned a doctoral degree, or left the country as of the survey reference date. The basic sampling design was a stratified design where strata were defined by 15 broad fields of study, 2 genders, and an 8-category "demographic group" variable combining race/ethnicity, disability status, and citizenship status. The 36,000 cases were to be allocated so that the nearly new cohort would have at least a 10 percent higher sampling rate than the old cohort to achieve increased accuracy of the estimates for the more recent graduates.

The sample cases were combined in the multiway cross of the stratification variables, which produced 240 strata. Proportional maintenance cut was used to determine the sample sizes across all 240 strata to maintain the sample size close to the sample sizes from 1999. In addition, a proportional cut from every stratum and selecting the sample using the probability proportion to size (PPS) selection technique decreased some of the within stratum weight variation. For strata where the allocated sample size was equal to the frame size, all cases were selected for the sample. For all other strata, sample cases were selected using the PPS selection method separately for each cohort group (with the sampling weights as the size measure).

The 2001 SDR new cohort sampling frame was created using the frame information from the 1999 and 2000 SED cases. Cases were considered ineligible for the sampling frame if they were missing a doctoral degree field from the SED, in a non-S&E doctorate field, in a post-doctorate location outside the United States, deceased, and age 76 or older as of the 2001 survey reference date. Stratified sampling design was also used for the new cohort sample to select approximately 4,000 cases from the sampling frame of 50,135 cases. This new cohort sample size ensured that the overall sampling rate for the new cohort would be at least 15 percent higher than the old cohort.

Because the sampling weight for every case in the new cohort sampling frame was equal to 1.0, a random method was the most appropriate method of sample selection. Each new cohort case was assigned a random number and was sorted by the random number. Each stratum had an allocated sample size for new cohort cases, and for any stratum where the allocated sample size was equal to the frame size, all cases were selected for the sample. For any other stratum, sample cases were randomly selected using the allocated sample size.

The overall sampling rate was about 1 in 17 (5.8 percent) in the 2001 SDR, applied to an estimated science and engineering doctoral population of 685,300. However, sampling rates varied considerably within and among the strata.

Survey Content

The 2001 SDR still retained the questionnaire design changes that were implemented in 1993. A large set of core data items is conveyed from year to year to enable trend comparisons. Each survey year, different sets of module questions on special topics of interest are included. For example, the 1995 SDR questionnaire had a postdoc module and the 1997 had special modules on alternative work arrangement, job security concerns, and recent doctorates' initial career experiences. In 2001, a special module on publication and patenting first introduced in 1995 was fielded again for activities during the past 5-year period. New questions were added in 2001 on individual satisfaction and importance of various job attributes. The multiple-race question, as mandated by the Office of Management and Budget (OMB), was also added in 2001 for research evaluation and future sampling purposes.

Data Collection

The 2001 SDR data collection consisted of two phases: a self-administered mail survey, followed by computer assisted telephone interviewing (CATI) of a sample of the nonrespondents to the mail survey. The mail survey consisted of an advance letter and then two mailings of a personalized questionnaire package, with a reminder postcard between the first and second questionnaire mailing.

In 2001, the SDR was mailed in two waves. Wave 1 consisted of the old (and nearly new) cohort component of the sample and Wave 2 consisted of the new cohort component. The data collection began for the old cohort

in April 2001 and for the new cohort in June 2001. The mailout process was the same for both cohorts.

The persons in the sample received a personalized advance letter from the Director of NSF to acquaint them with the survey. A week later, the first mailout questionnaire was sent out, followed by a reminder/thank you postcard the following week. Approximately 7 weeks after the first mailout, the sample members who did not return a completed questionnaire from the first mailout received a second questionnaire mailing via USPS Priority Mail.

Prior to mailout, name and address information of the sample members was updated using a variety of resources, such as FastData[SM] and PostalSoft[TM]. Extensive quality control measures were conducted to ensure accurate mailing information before the first mailing. Address updates from the first mailout were made prior to the second mailing. Other cases went through an extensive locating process prior to CATI and continued until end of the data collection.

The CATI follow-up of mail nonrespondents began in August 2001 for the old cohort and in September for the new cohort. The CATI was completed for all cohorts in November 2001.

Response Rates

The overall unweighted response rate for the 2001 SDR was 82.2 percent. The response to the mail phase of the survey was 66.8 percent. The response rate to the CATI phase was 46.9 percent. The overall weighted response rate was 82.6 percent (weighted response divided by the weighted sample cases). The response rates for the new cohort and old cohort were 83.3 percent and 82 percent, respectively. Among the old cohort, about 92.0 percent of the respondents and 36.8 percent of the nonrespondents in 1999 responded to the 2001 survey.

Data Preparation

Data preparation for the 2001 SDR consisted of clerical, keying, and coding operations performed manually by the Census National Processing Center (NPC) and the computer operations performed by the Census Demographic Surveys Division (DSD). Data preparation began in May 2001 when the first mail questionnaires were returned to the NPC and continued through August 2002 when the DSD delivered the SESTAT formatted, edited, and imputed data file to the NSF.

As the mail questionnaires were received, they were checked into the tracking system. The mail-returned questionnaires that had one or more entries were clerically edited for data entry preparation. The clerical edit was limited to simple edits such as correcting illegible entries, rounding fractions to the closest whole number, verifying that city, state, and country entries were in the correct location.

Clerically edited questionnaires were grouped into batches, keyed, and verified using the Key Entry III (KE III) system. The KE III system generated a keying report to track the status of cases through the keying operation. As part of quality control procedures, 5 percent verification was performed of all keyed questionnaires. For some questionnaire items (F12 birthdate, F16/F17/F19 contact information), a 100 percent verification of questionnaire items was performed.

NPC transmitted the keyed questionnaire data on a regular basis during the data collection phase to the DSD. DSD performed computer editing to identify cases with missing critical items (A1/A2 labor force status, A6/A21 job codes, F4 resident status in U.S., F12 birthdate) and generated telephone follow-up sheets. Telephone callbacks were made to obtain responses to these critical items; otherwise they were considered incomplete responses. Whenever these callbacks were made, every attempt was also made to obtain responses to other missing important data items (A7 full/part-time status, A15/A17 type of employer, A18/A19 faculty rank and tenure, A26 job start date, A30/A31 work activities, and F17 future contact information). Overall, about 7 percent of the completed mail respondents required a telephone followup for responses to the missing critical items.

Because the DSD collected data in mail and CATI, the data sets were merged into one data set. The coding operation involved special coding of *occupation and education codes*, *other specify coding, state and country coding,* and *IPEDS coding*. For special coding of occupation, the respondent's occupational data were reviewed along with other work-related data from the questionnaire by specially trained coders to "correct" known respondent self-reporting problems to obtain the "best" occupation codes. The education code for a newly earned degree was assigned strictly based on the degree field verbatim.

The "Other Specify" responses were back-coded to existing response categories using the SESTAT other specify coding guidelines. Employer location (A11),

Degreed school location (D6) and Country of citizenship (F8) were assigned the appropriate three-digit FIPS state/country code. The Integrated Postsecondary Education Data System (IPEDS) was used to assign codes for the employers (A11) that are postsecondary institutions and for the newly earned degree school (D6).

A detailed edit specification was developed from the SESTAT edit guidelines to perform further computer editing of multiple values to "Mark One" questions, skip errors, range errors, interitem inconsistencies, cross-year inconsistencies. Basic frequency distributions of all survey items showed item nonresponse rates to be generally less than 3 percent. Nonresponse to a few questions deemed somewhat sensitive, such as annual salary, were around 5.4 percent.

To compensate for item nonresponse, data not reported by the respondents as well as responses of "refused" or "don't know" were imputed. Imputation is a process for treating missing data. Imputation methods are used when answers to questions are blank or not usable. Two imputation methods were used: (1) logical imputation, and (2) hot deck imputation. For logical imputation, either the respondent's answers to related questions determined what the missing value had to be, or the respondent's answer to the same question in the prior survey round was substituted for the missing value. The latter approach of using the historical data is often called "cold deck" imputation. Cold deck imputation is useful for variables that are static, such as place of birth or gender. When logical imputation was used, it was employed before hot deck imputation.

In hot deck imputation, a donor case is selected from the current round of respondents by matching on related variables. The donor case's response is used as a proxy for the recipient's missing variable. Hot deck imputation is the method of choice for variables that may change over time, such as employment characteristics. Hot deck is preferable to model-based imputation in this application because it easily preserves correlation among variables and maintains the valid response ranges for categorical variables.

Weighting and Estimation

To enable weighted analyses of the 2001 SDR data, a sample weight was calculated for every person in the sample. The primary purpose of the weights is to create representative estimates by adjusting for unequal probabilities of selection. The second purpose is to adjust for the effects of nonresponse without increasing the variance. Informally, a sampling weight approximates the number of persons in the doctorate population that a sampled person represents. A main goal of this weighting plan is to produce final weights that reduce the nonresponse bias in our survey estimates, without increasing the variance.

The weights were calculated in several stages. The first stage was the calculation of base weights that account for the sample design. A base weight is the inverse of the probability of selection into the SDR sample. For cases selected with certainty, the 2001 SDR base weight is equal to the 2001 SDR initial weight. For all other cases, the 2001 SDR base weight is greater than the initial weight. This increase reflects an adjustment for cases not selected for the sample.

From the 2001 SDR base weights, the production of the 2001 SDR final weights involved four main steps:

- Adjustment for duplicate, frame ineligible, and never earned doctorate cases
- Calculation of the 2001 SDR control totals
- Calculation of the 2001 SDR noninterview weights
- Calculation of the 2001 SDR final weights

Raking ratio adjustment was used to control the 2001 SDR sample back to the 2001 SDR population totals. The purpose of this adjustment is twofold:

- To decrease the sampling variability
- To account for changes in the final weights due to changes in the eligible sampling frame

Reliability

Because the estimates produced from this survey are based on a sample, they may vary from those that would have been obtained if all members of the target population had been surveyed (using the same questionnaire and data collection methods). Two types of error are possible when population estimates are derived from any sample survey: sampling error and nonsampling error. By looking at these errors, it is possible to estimate the accuracy and precision of the survey results.

Sampling error is the variation that occurs by chance because a sample, rather than the entire population, is surveyed. The particular sample that was used to estimate the 2001 population of science and engineering doctorates in the United States was one of a large number of samples

that could have been selected using the same sample design and size. Estimates based on each of these samples would have differed. Thus, one should be particularly careful when interpreting results based on a relatively small number of cases or on small differences between the estimates.

Due to the large amount of data collected in the SDR, it is not practical to directly calculate variance estimates for every survey estimate. Instead, generalized variance functions were developed to model the variance estimates for certain characteristics. Parameters derived from these generalized variance functions approximate variance estimates for all survey items. As a result, these sampling errors provide an indication of the order of magnitude of a sampling error rather than a precise sampling error for any specific item.

The variances on the survey estimates were calculated by the successive difference replication method. This replication method was used to first calculate a small number of variance estimates, which were then used to estimate the parameters of the generalized variance function. A one-parameter model was used to calculate the generalized variance parameters which were estimated using an iterative weighted least square procedure.

Since many of the SDR estimates of interest consist of small populations such as estimates of Hispanic scientists or black engineers, the finite population correction factor was consistently applied to all the variance estimates.

Different generalized variance functions were used to estimate standard errors associated with a broader range of totals and percentages. The *a* and *b* parameters were calculated for each of the demographic groups and fields of study shown in Appendix C. The *a* and *b* parameters can be used to approximate standard errors for the S&E doctoral population overall, for broad field groupings used by NSF, and for selected subgroups of analytic interest.

Standard Error of Estimated Numbers

To calculate the desired standard errors on numbers, let X denote the estimated number. The standard error can be approximated using the appropriate values of a and b along with the following formula for standard errors of totals:

$$SE(X) = [aX^2 + bX]^{1/2} \quad (1)$$

When calculating standard errors for numbers from tabulations involving different characteristics, use the set of parameters for the characteristic that will give the largest standard error.

Illustration

Suppose an estimated 3,240 females with doctorates in the biological sciences were reported as working in the Federal Government in 2001.

Use the appropriate generalized variance parameters from Appendix C to get:

- Survey estimate X = 3,240
- a parameter = -0.000079
- b parameter = 13.0148

Use formula (1) to approximate the standard error on the estimated number of 3,240 as:

$$SE(X) = [(-0.000079 \times 3{,}240^2) + (13.0148 \times 3{,}240)]^{1/2} = 203$$

The 95% confidence interval is calculated using the following formula:

$$95\%\ CI = X \pm [1.96 \times SE(X)] \quad (2)$$

where

X is the survey estimate of interest, and $SE(X)$ is the estimated standard error for the survey estimate of interest.

Using formula (2) above, the 95% confidence interval is:

$$3{,}240 \pm 1.96 \times 203 \text{ or } 3{,}240 \pm 398$$

Therefore, the 95% confidence interval has the following limits:

- Lower limit = 2,842
- Upper limit = 3,638

So we can say with 95% confidence that the number of females with biological sciences doctorates working in the Federal Government in 2001 is estimated to be between 2,842 and 3,638.

Standard Error of Estimated Percentages

To calculate the standard errors on percentages, let p equal the percentage possessing the specific characteristic and X and Y represent the numerator and denominator, respectively, of the ratio that yields the observed percentage. The standard error of a percentage may be approximated using the formula:

$$SE(p) = p\{[([SE(X)^2)/X^2] - ([SE(Y))^2/Y^2]\}^{1/2} \quad (3)$$

where

X and Y are survey estimates of interest, SE(X) and SE(Y) are the corresponding standard error estimates derived using formula (1), and p is the estimated percentage ($p = X/Y \times 100$).

Illustration

Suppose an estimated 3,240 of the 11,530 biological sciences doctorates working in the Federal Government are women. Therefore, the estimated percentage of biological sciences doctorates working in the Federal Government who are women is 28.1%.

Use formula (1) and the appropriate parameters from Appendix C, to get:

	X	***Y***	***p***
• Survey estimate	3,240	11,530	28.1%
• *a* parameter	-0.000079	-0.000114	NA
• *b* parameter	13.0148	18.7606	NA
• Standard error	203	448	

Insert the above numbers into formula (3) to approximate the standard error of the estimate of 28.1%:

$$SE(p) = 28.1\ [(203^2/3{,}240^2) - (448^2/11{,}530^2)]^{1/2} = 1.4\%$$

Using formula (2), the 95% confidence interval is:

$$28.1\% \pm 1.96 \times 1.4\% \text{ or } 31.2\% \pm 2.7\%$$

Therefore, the 95% confidence interval has the following limits:

- Lower limit = 28.5%
- Upper limit = 33.9%

Standard Error of a Difference

To calculate the standard errors of the difference between two sample estimates, let X and Y represent two estimates of interest and SE(X) and SE(Y) the corresponding standard error estimates derived using formula (1).

$$SE(X\text{-}Y) = \{[SE(X))^2 + (SE(Y)]^2\}^{1/2} \quad (4)$$

The estimates can be numbers, percentages, ratios, etc. This will represent the actual standard error quite accurately for the difference between estimates of the same characteristic in two different areas or for the difference between separate and uncorrelated characteristics in the same area.

Illustration

In 2001, suppose there were an estimated 8,290 male and 3,240 female biological sciences doctorates working in Federal Government. The apparent difference between the estimated number of male and female biological sciences doctorates is 5,050.

Use the appropriate parameters from Appendix C and formula (1) to get:

	X	***Y***	**Difference**
• Survey estimate	8,290	3,240	5,050
• *a* parameter	-0.000114	-0.000079	NA
• *b* parameter	18.7606	13.0148	NA
• Standard Error	448	203	

The standard error of the difference is calculated using formula (4):

$$SE(X\text{-}Y) = (448^2 + 203^2)^{1/2} = 497$$

The 95% confidence interval is calculated as 5,050 $\pm$ 1.96 $\times$ 497 or 3,330 $\pm$ 974. Because this interval does not include zero, we can conclude with 95% confidence that the estimated number of male biological sciences

doctoral recipients working in Federal Government is significantly higher than the number of female biological sciences doctoral recipients.

However, if there is a high positive/negative correlation between the two characteristics, the formula will overestimate/underestimate the true standard error.

In addition to sampling error, data are subject to nonsampling error, which can arise at many points in the survey process. Sources of nonsampling error take many different forms: (1) nonresponse bias, which arises when the characteristics of individuals who do not respond to a survey differ significantly from those who do; (2) measurement error, which arises when we are not able to precisely measure the variables of interest; (3) coverage error, which arises when some members of the target population are not identified and thus do not have a chance to be selected for the sample; and (4) processing error, which can arise at the point of data editing, coding, or key entry. These sources of error are much harder to estimate than sampling errors.

Important Notes on the Tables

The following definitions are provided to help facilitate the use of data in the detailed tables.

Field of doctorate is the field of degree as specified by the respondent in the Survey of Earned Doctorates (SED) at the time of degree conferral. These codes were subsequently recoded to the SESTAT codes. (See Appendix A for the doctorate degree fields.)

Occupation data were derived from responses to several questions on the type of work primarily performed by the respondent. The occupational classification of the respondent was based on his/her principal job held during the reference week—or last job held, if not employed in the reference week (questions A20 or A5). Also used in the occupational classification was a respondent-selected job code (questions A21 or A6). (See Appendix B for the list of occupations.)

Sector of employment was based on responses to questions A15 and A17. The category "universities and 4-year colleges" includes 4-year colleges or universities, medical schools (including university-affiliated hospitals or medical centers), university-affiliated research institutions, and other types of educational institutions. "Private-for-profit" includes those self-employed in incorporated business.

Employer location was based primarily on responses to question A11 on the location of the principal employer. Individuals not reporting place of employment were classified by their last mailing addresses.

Primary work activity was determined from responses to question A31. "Development" includes the development of equipment, products, and systems. "Design" includes the design of equipment, processes, and models.

Federal support was determined from responses to questions A42 and A43.

Faculty rank/tenure status was obtained from the responses to questions A18 and A19.

Race/ethnicity categories of white, black, Asian/Pacific Islander and American Indian/Alaskan Native refer to non-Hispanic individuals only. These data are from the SED.

Citizenship status category of non-U.S., temporary resident does not include individuals who at the time they received their doctorate reported plans to leave the U.S. These individuals were excluded from the sampling frame.

Salary data were derived from responses to question A35, in which information was requested regarding annual salary before deductions for the principal job held during April 2001, excluding income from bonuses, overtime, and summer teaching/research. Salaries reported are median annual salaries, rounded to the nearest $100 and computed for full-time employed scientists and engineers. For individuals employed by education institutions, no accommodation was made to convert academic-year salaries to calendar-year salaries. Users are advised that due to changes in the salary question since 1993, the 1995 through 2001 salary data are not strictly comparable with the 1993 salary data.

Labor force participation rate. The labor force is defined as those employed (E) plus those unemployed (U, those not-employed persons actively seeking work). Population (P) is defined as all S&E doctorate holders under age 76, residing in the United States during the week of April 15, 2001, who earned their doctorates from U.S. institutions. The labor force participation rate (R_{LF}) is the ratio of the labor force to the population (P).

$R_{LF} = (E + U) / P$

Unemployment rate. The unemployment rate (R_u) is the ratio of those who are unemployed (U) to the total labor force (E + U). $R_u = U / (E + U)$

Involuntarily out-of-field rate. The involuntarily out-of-field rate is the percent of employed individuals who reported they were either:

- Working part-time exclusively because suitable full-time work was not available and/or
- Working in an area not related to the first doctoral degree (in their principal job) at least partially because suitable work in the field was not available.

Summary of Table Changes in 2001 Compared to 1999 Tables

Global changes

1. Occupation tables now show separate groups of the postsecondary teachers wherever possible.
2. In all occupation tables, "S&T Historians and other social scientists" job category title was changed to "Other social scientists."

Specific table modifications in 2001

Table 1 Postdoc column was removed from the table and the postdoc estimates by doctorate field are separately reported in new table 7.

Table 25 Postdoc column was removed from the table and the postdoc estimates by occupation are separately reported in new table 31.

New Tables in 2001

Table 7
Table 8
Table 31